Twitter Event Signatures: An investigation of how real-world events are reflected in Twitter signals.

Natan Grayman

School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

Abstract: The aim of this investigation report is to elucidate the process of exploring the intricate relationship between various event types and the corresponding patterns of interest and trends observed on Twitter. Extensive research into open-source tweet databases was undertaken, resulting in the use of datasets from platforms including Kaggle, the Internet Archive, and GDELT. The event types analysed encompassed protests, natural disasters, and election cycles, each of which involved the collection and extraction of two sets of event datasets. The analysis of each compiled dataset extended to various dimensions, including geolocation, news media response, and trend validation. Mathematical techniques such as seasonal decomposition and the LOESS regression method were employed to approximate and evaluate the trends. Ultimately, this facilitated the creation of a framework for classifying each event with a distinct event archetype. The results and key findings from this investigation offer valuable insights into the relationship between event types and Twitter trends, paving the way for more precise event classification and trend analysis.

Key words: *Twitter trends, event archetype, open-source tweet data, seasonal decomposition, LOESS regression.*

INTRODUCTION

In this investigation project, the primary objective is to explore the intricate relationship between various event types and the corresponding patterns of interest and trends observed on Twitter. Through this investigation the overarching question arises of how real-world event types are reflected in the nuanced signals of social media, particularly Twitter. This analysis and comprehension of social media responses to different event types can yield novel insights into online interactions and the foundations of contemporary society [1].

The following sections will guide through the key aspects of the investigation. Commencing with a background section, the concept of event signatures is explored, elucidating their manifestations and examination on Twitter, along with background of other open-source platforms. Subsequently, the investigation methodology explains the approach to restrict and define the event types under examination, as well as the methods utilized to retrieve the corresponding data.

Thereafter, the key findings and results section provides a detailed explanation of the results at each stage of the investigation project, offering insights into the discovered patterns and trends. Finally, the discussion section critically assesses the project's limitations and outlines future research prospects.

1. BACKGROUND

In this section the background on the literature on what is an event signature and the impact of the change in ownership of Twitter on this investigation project will be discussed.

1.1. Event Signatures

In the digital age, the pervasive influence of social media platforms has reshaped the way people interact with live events, enabling users to instantaneously engage, react, and participate in global discussions. Consequently, during significant societal occurrences, be they natural disasters or political upheavals, individuals often turn to these platforms as their initial response, forming a digital footprint that chronicles their reactions [1]. This digital footprint can be encapsulated by the term "Event Signatures," signifying the distinctive patterns of public interest that emerge in response to various types of events [1].

The “Event Signature' is a comprehensive representation of any reactive metric, depicting the evolving user engagement and interest in a particular event [2]. This comprehensive digital footprint encompasses user activities, such as posts, comments, likes, edits, reposts, and hashtag usage, collectively reflecting the degree and nature of public interest in the given event [2].

The analysis of event signatures offers invaluable insights into the inherent qualities of the events, facilitating the development of future mitigation and response strategies [1] [2].

1.2. Twitter

Twitter, the prominent microblogging platform known for its real-time information sharing, underwent substantial changes in 2023 following its acquisition by Elon Musk [3]. These changes included a rebranding to 'X' and the monetization of its API [3].

By analysing event signatures on Twitter, it becomes

possible to examine the dynamic and temporal volume of tweet reactions to a particular event, revealing the pattern of interest over time [1] [2]. This analysis is facilitated by extracting tweets based on keywords, hashtags, and dates [3]. The event signature, in this context, can be described as a numeric sequence representing the daily tweet volume:

$$S = \{s_1, s_2, \dots, s_n\} \quad (1)$$

Where:

s_i = the volume of event-related content in time period i.

1.3. GDELT

Supported by Google, the Global Database of Events, Language, and Tone (GDELT) is a dynamic and continuously updated dataset that comprehensively monitors and records a wide array of global events, news articles, and media sources worldwide *[4]*. GDELT is recognized as the largest open-source database of human society, making it an invaluable resource for researchers, analysts, and data scientists seeking insights into global events, trends, and sentiments spanning several decades *[4]*.

2. INVESTIGATION METHODOLOGY

This section delves into the investigation methodology of data collection and its impact on the selection of three distinct event types for investigation.

2.1. Data Collection

Despite the constraints posed by the monetization of the Twitter API, the investigation project proceeded and overcame the issue by utilizing open-source data. Consequently, the methodology for selecting events to investigate was contingent on the availability of high-quality tweet datasets corresponding to specific event types.

In the early stages of dataset research, the Wharton and Annenberg Historical Dataset appeared promising for high-quality data [5]. Unfortunately, access to this dataset was denied. Subsequent exploration of various GitHub repositories revealed issues such as data corruption, insufficient volume, and dataset inconsistencies.

Nevertheless, the investigation resulted in the identification of a substantial data source archive known as The Internet Archive. This 501(c)(3) non-profit organization is dedicated to creating a digital library of internet sites and other cultural artifacts in digital form [6].

This dataset proved to be extensive, with an impressive volume of approximately 4,000 to 5,000 tweets recorded every minute within the specified date range [6]. However, it is important to note two specific constraints associated with this dataset. Firstly, the dataset's temporal coverage is restricted, as it does not include tweets from November 2022 to September 2023, thereby limiting the analysis of events to the earlier time frame. Secondly, the tweet data extraction method is a light tier, lacking tweet features such as longitude and latitude information, a deficiency discussed in Section 3.2.

Throughout the investigation project, data extraction from the Internet Archive was effectively refined and optimized. Initially, the archive's data required decompression and was searched using a variety of relevant keywords, hashtags, and locations. This extraction process was both tedious and error-prone, leading to errors in dataset formatting and extraction handling.

Likewise, the initial use of a local computer for extraction resulted in slow processing and high battery consumption. However, these initial challenges provided valuable learning opportunities. Subsequently, the team transitioned to establishing efficient access and usage of the Wits cluster while implementing multiprocessing for extraction. This optimized process enabled the utilization of a template code for various events, streamlining the addition of new datasets for archetype creation.

Furthermore, Kaggle, an online community platform for data scientists and machine learning enthusiasts, served as another crucial source for data collection [7]. The utilization and leveraging of high-quality data on Kaggle, combined with the flexibility of the Internet Archive, enabled the team to make well-informed decisions regarding which event types to investigate.

2.2. Event Selection

The event types selected based on the availability of ample, high-quality data were Elections, Natural Disasters, and Protests. Specifically , the events included the 2020 US election tweets from Kaggle, the 2023 Turkey and Syria Earthquake tweets from Kaggle, and the 2021 South African unrest tweets extracted from the Internet Archive.

Following the methodical extraction, cleaning, and exploration of these datasets, the investigation into signature analysis was initiated, facilitating an in-depth exploration of the distinctive features within each event's signature.

3. RESULTS AND KEY FINDINGS

To ensure comprehensive access to all the results, the project team has generated three Kaggle Jupyter notebooks, utilized for its cloud computation capabilities and shared workspace. These notebooks cover each event type: Investigation of Protest Twitter Event Type, Investigation of Election Event Type, and Investigation of Natural Disaster Event Type.

Each Jupyter notebook adheres to a similar structure,

offering a table of contents and detailed explanations. Nonetheless, this section will highlight the key findings from the notebooks, providing an overview of the investigation findings that ultimately led to the event archetype results.

3.1. Pattern of Interest

The primary goal of this investigation project, which aimed to uncover the pattern of interest for each distinct event type, was successfully achieved. Appendix C illustrates the pattern of interests identified for each of the initial events within the Protest, Election and Natural Disaster event type with contextual date markings.

When comparing the patterns of interest, it became evident that the varying count sizes within each dataset rendered graph comparisons challenging due to their distinct scales. To address this issue, normalization, a statistical technique employed to rescale data and bring it within a common range, was utilized [8]. It proved crucial throughout the investigation project in facilitating a fair assessment of the shapes and trends of different curves, preventing variations in magnitude or scale from overshadowing the underlying structural patterns.

3.2. Geolocation

Analysing the features of each Twitter dataset revealed key findings that deepened the understanding of the data and directed subsequent steps in the investigation. One notably interesting aspect was Twitter's ability to provide geotagging coordinates in the form of longitude and latitude features. This facilitated the creation of geographical maps for each event. However, it's important to note that this feature was exclusively available in Kaggle datasets containing geotagging information. In contrast, the Internet Archive dataset provided solely textual location descriptions gathered by Twitter. To convert these textual location descriptions into geotagging coordinates, the open-source Nominatim API was employed [9]. Figure 1 illustrates the geocoded map of reactions during the 2021 South African Social Unrest.

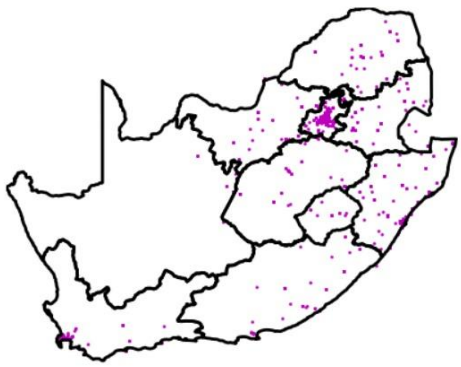

Figure 1:Geolocation map of Twitter activity during the 2021 South African Social Unrest.

While addressing the challenge of missing coordinate features, this exploration led to an examination of GDELT's functionality.

3.3. News Comparison

Through GDELT's open-source data, accessible via Google BigQuery, the total number of news articles published by registered news companies during each event can be extracted. This enables a visual comparison between the shape and trend of the news coverage and social media response. Appendix D illustrates this comparison for each event type.

Using Figure 8 in Appendix D of the South African social unrest as an example, a noticeable delay can be observed, with the news media peaking in response later compared to the peak in social media activity, notably when the army was deployed.

3.4. Trend Validation

Investigating both the pattern of interest in tweets and the temporal changes in news articles over the same duration raised questions about the validity of the observed trends. This led to an exploration of additional sources, including the Mastodon API, an open source, decentralized social media platform similar to Twitter [10].

After extracting Mastodon's posts, known as 'toots,' it became evident that there was insufficient data available for both the South African Social Unrest and the Turkey-Syria unrest events [10]. Nevertheless, the extraction for the US election yielded ample data, allowing validation of the observed trend during the election cycle.

Additionally, Google Trends played a crucial role in the trend validation process. Google Trends is a website that analyses the popularity of search queries in Google Search across various regions and languages, providing users with graphical trends to compare the search volume of different queries [11]. It also features clustered topics related to trending searches on Search, Google News, and YouTube, offering insights into both live and historical search trends that capture people's reactions [11]. Therefore, utilizing the trend functionality for search volume trend data for both search queries and news articles allowed the validation of the general trend of user reactions throughout each event's duration.

Using Figure 11 in Appendix E , evidently, the trend observed on Google by users searching for key terms like "Looting," "State of Emergency," and "Protest" closely follows the upward peak and gradual downward trajectory of the pattern of interest in Figure 5.

3.5. Mathematical Framework

To investigate the trends observed in all events, a

mathematical trend analysis was conducted. Seasonal decomposition emerged as an effective framework for analysing the time-series data [8]. Seasonal decomposition entails breaking down the time-series data into three primary components [8]. The trend component reveals the underlying long-term behaviour within the data, shedding light on whether it displays a general trend of increase, decrease, or stability over time [8]. Seasonal patterns, characterized by recurring fluctuations at fixed intervals, are captured within the seasonal component, aiding in the identification of consistent patterns in events or data [8]. The residual component encompasses unexplained or random fluctuations, proving valuable for detecting anomalies or noise not explicable by trends or seasonality [8].

Furthermore, seasonal decomposition provides two core methods, additive and multiplicative, to break down data into trend, seasonal, and residual components, with the additive approach applicable when seasonal effects remain constant over time and the multiplicative approach suitable when the seasonal influence scales with the data's level [8]. Moreover, a "period" parameter in seasonal decomposition defines the number of data points within a single season, enabling alignment with the data's inherent seasonal patterns. As a result, the exploration of seasonal decomposition involved the use of a slider widget to explore various periods and a widget to choose between additive or multiplicative methods [8].

The resulting seasonal decomposition verified the hypothesis that a small period , of a few days within each event , showed an effective aligned trend, daily seasonal fluctuation and a relatively low residual curve. However, the results obtained from seasonal decomposition are limited by their reliance on trend curves that seek to identify long-term general patterns of growth, decline, or stability over time, rather than capturing the inherent complexity within each event type.

Therefore, to utilize a more granular and controlled technique for approximating the trend of each pattern of interest, the LOESS (Locally Estimated Scatterplot Smoothing) non-parametric regression technique was employed [12]. This approach offers the advantage of a smoothing parameter that can be adjusted through trial and error to effectively increase or decrease the smooth bandwidth for each pattern of interest [12]. However, despite LOESS's flexibility in modelling complex curves for which no theoretical models exist, it does not produce a regression function that is easily represented by a mathematical formula [12].

3.6. Event Archetype

After analysing each event, including its pattern of interest, news comparison, trend validation, and mathematical framework, the investigation progressed to explore additional data sources for inclusion. Thanks to the optimized data extraction processes, lessons learned throughout this study and streamlined code templates, the seamless integration of new datasets into each event type was facilitated.

In the Protest event type, investigations were conducted to consider the inclusion of both the Black Lives Matter Protests in 2020 and the Mahsa Amini Iran Protests. However, the Black Lives Matter tweet data examined was found to lack sufficient data for the entire event. In the Election event type, the US election cycle in 2016 was introduced. For the Natural Disaster event type, the Mexico earthquake in September 2017 was integrated. Following the incorporation of these events into the Jupyter notebook for each event type, the LOESS mathematical framework was employed to explore each trend and ultimately examine the feasibility of formulating an archetype for each event type.

The resulting archetypes, established with the use of upper and lower boundaries, represent the potential range within which the Twitter trend could be positioned for each event type. The implementation of these boundaries is essential for capturing the average trend between the curves, offering a range in which the archetype is situated and allowing for a more accurate representation of the typical behaviour of each event type.

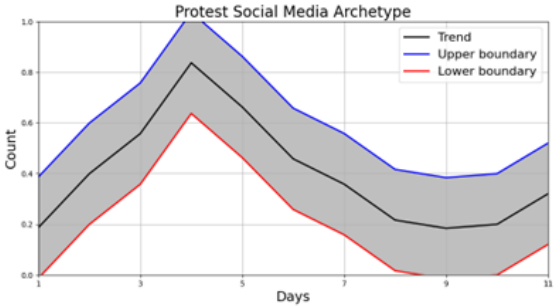

Figure 2: Event Archetype of Protest event type

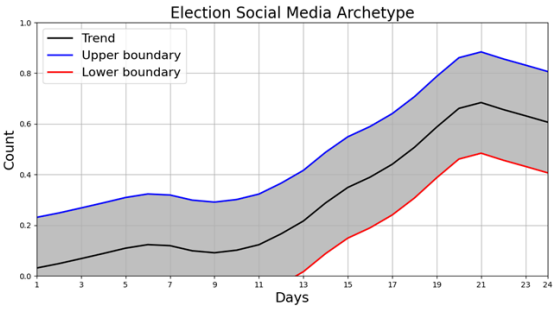

Figure 3: Event Archetype of Election Archetype

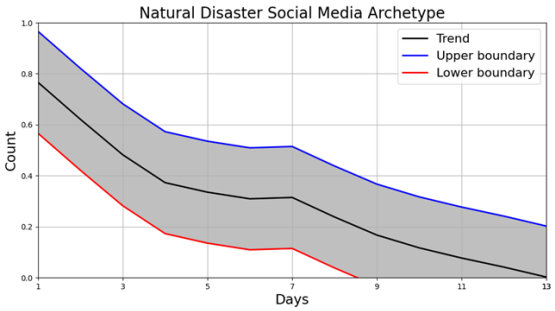

Figure 4: Event Archetype of Natural Disaster

To validate the resulting archetypes, similarity metrics were employed to assess the proximity of each pattern of interest to the overall archetype. The first metric employed is the Euclidean distance, which quantifies the spatial or geometric distance between two data points [13]. Since both curves are normalized, the Euclidean distance is suitable for measuring the magnitude of points or outliers from the archetype trend [13].

Likewise, the Mean Squared Error (MSE) metric was employed to assess the average squared differences between two datasets, offering a measure of their overall dissimilarity [14]. This metric proves suitable for comparing the archetype to the pattern of interest as it enables the evaluation of error magnitudes in trend data, facilitating a meaningful assessment of their overall fit [14]. The range of MSE values enables the quantification of how closely the pattern of interest aligns with the archetype, with lower MSE values signifying a closer match and higher values indicating greater dissimilarity [14].

Furthermore, the Cosine Similarity metric was employed to facilitate directional relationship quantification [15]. It calculates the cosine of the angle between the two compared data points and, as a result, quantifies the directional similarity between the points in each curve [15]. The scale of Cosine Similarity ranges from -1, signifying complete dissimilarity, to 1, indicating perfect similarity, while a value of 0 suggests that the data points are unrelated or uncorrelated [15]. Similarly, to explore the linear directional relationship between the archetype and each pattern of interest, the Pearson correlation coefficient was employed [16]. To illustrate these results, Table 1 shows the resulting similarity metrics for the Election event type.

Another mathematical feature investigated was the use of cross-correlation to quantitatively model the time delay between social media and news media responses. By treating the response curves as signals and employing convolution, cross-correlation identifies the time lag at which the two signals exhibit maximum similarity or divergence [17]. The cross-correlation formula for two discrete datasets X and Y, each with n data points, is expressed as:

$$R(k) = \sum(X(t) \cdot Y(t-k)) \quad (2)$$

Where:
R(k) = denotes the cross-correlation at time lag k [17].

Table 1 showing the similarity metrics comparing the US election pattern of interests to the event archetype.

Pattern of Interest	Euclidean Distance	Cosine Similarity	Pearson Correlation	Mean Squared Error
US Election 2020	1.2998	0.8759	0.8177	0.0298
US Election 2016	0.5350	0.9762	0.9487	0.0309

4. DISCUSSION

4.1. Limitations

The investigation project faced an initial setback with Twitter's ownership structure changes, leading to the inaccessibility of the Twitter/X's API as a data source, primarily due to the substantial financial allocation required, given the extensive data volume involved in this project (exceeding a terabyte!). Consequently, the utilization of open-source Twitter datasets in this research introduced several notable limitations. Firstly, these datasets could exhibit inherent biases that could predominantly reflect specific demographics or characteristics of Twitter users, potentially constraining the generalizability of findings. Moreover, the presence of potential data fabrication or bot-generated content within open-source Twitter datasets could raise concerns about the authenticity and reliability of the information.

These limitations also extended to the scope of the data, as open-source datasets, such as the Internet Archive, often offered only a partial representation of the broader Twitter landscape, risking the omission of crucial tweets or trends that could significantly impact research outcomes.

Similarly, even though a well-defined methodology was meticulously employed, encompassing the use of popular keywords and hashtags for the systematic data extraction from sources like the Internet Archive and GDELT, the inherent risk of incomplete data retrieval persisted, thereby giving rise to potential gaps within the dataset. Despite the rigorous approach, certain factors such as variations in keyword usage or the occasional absence of relevant hashtags could lead to unaccounted-for portions of data. These limitations imply that the datasets utilized might not encapsulate the entirety of conversations and discussions relevant to the chosen event types, limiting the comprehensiveness of the analysis.

4.2. Future Research

The outcomes of this investigation project have laid a robust foundation for potential future research endeavours. The establishment of a comprehensive template for extracting tweets from reputable Twitter archives, exemplified by the Internet Archive, along with the subsequent procedures for data analysis ,cleansing and mathematical formulations, has not only streamlined the investigative process but also exemplified a heuristic approach that can be replicated and extended to encompass additional event types. As the methodology becomes further refined and event data accumulates, this template may serve as a valuable tool, analogous to a 'sensor,' enabling real-time Twitter monitoring and event forecasting, specifically in the context of protests, natural disasters, or election cycles. By leveraging the event archetype, future research can delve into forecasting mechanisms and user interaction analysis with a heightened level of precision and accuracy, offering novel insights into these critical aspects of contemporary society.

5. CONCLUSION

In conclusion, this investigation aimed to shed light on the process of exploring the intricate relationship between diverse event types and the corresponding patterns of interest and trends observed on Twitter. Extensive research was conducted, leveraging open-source tweet databases from platforms such as Kaggle, the Internet Archive, and GDELT. The study focused on three key event types: protests, natural disasters, and election cycles, each involving the collection and extraction of two sets of event datasets. The comprehensive analysis of each compiled dataset extended across multiple dimensions, including geolocation, news media response, and trend validation. To approximate and assess the trends within these datasets, mathematical techniques like seasonal decomposition and the LOESS regression method were thoughtfully employed. The culmination of these efforts resulted in the creation of a framework for classifying each event with a distinct event archetype. The invaluable insights gleaned from the results and key findings in this investigation offer a deeper understanding of the intricate relationship between event types and Twitter trends, setting the stage for more precise event classification and trend analysis in the future.

REFERENCES

[1] Y. X. ,. D. R. Haji Mohammad Saleem, “Effects of disaster characteristics on Twitter event signature.,” ScienceDirect, Montreal, 2014.

[2] J. L. Jaewon Yang, “Patterns of temporal variation in online media,” WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining. ACM Digital Library, 2011.

[3] X Developer Platform, “X API,” 2023. [Online]. Available: https://developer.twitter.com/en/products/twitter-api . [Accessed 23 October 2023].

[4] GDELT, “The GDELT Project,” [Online]. Available: https://www.gdeltproject.org/about.html. [Accessed 24 October 2023].

[5] Wharton University of Pennsylvania, “Dataset of Historical Tweets,” [Online]. Available: https://research-it.wharton.upenn.edu/data/tweet-database/. [Accessed 23 October 2023].

[6] Internet Archive, “About the Internet Archive,” [Online]. Available: https://archive.org/about/. [Accessed 23 October 2023].

[7] Kaggle, “Datasets,” [Online]. Available: https://www.kaggle.com/datasets. [Accessed 23 October 2023].

[8] G. A. Rob J Hyndman, Forecasting: Principles and Practice (2nd ed), Monash University, Australia: Otexts, 2023, pp. 156-238.

[9] Nominatim Manual, “Overview,” Nominatim developer community, 2023. [Online]. Available: https://nominatim.org/release-docs/develop/api/Overview/. [Accessed 24 October 2023].

[10] Mastodon, “What is Mastodon?,” [Online]. Available: https://docs.joinmastodon.org/. [Accessed 24 October 2023].

[11] Google, “Google Trends,” [Online]. Available: https://trends.google.com/trends/. [Accessed 24 October 2023].

[12] W. S. Cleveland, “LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression,” The American Statistician, Vol. 35, No. 1 (Feb., 1981), p. 54 (1 page), 1981.

[13] NIST, “Euclidean distance,” [Online]. Available: https://xlinux.nist.gov/dads/HTML/euclidndstnc.html . [Accessed 25 October 2023].

[14] Britannica, “mean squared error,” [Online]. Available: https://britannica.com/science/mean-squared-error. [Accessed 25 October 2023].

[15] F. Karabiber, “Cosine Similarity,” LearnDataSci, [Online]. Available: https://www.learndatasci.com/glossary/cosine-similarity/. [Accessed 25 October 2023].

[16] KentState University , “SPSS TUTORIALS: PEARSON CORRELATION,” [Online]. Available: https://libguides.library.kent.edu/spss/pearsoncorr. [Accessed 25 October 2023].

[17] All About Circuits, “Understanding Correlation,” [Online]. Available: https://www.allaboutcircuits.com/technical-articles/understanding-correlation/. [Accessed 25 October 2023].

Appendix A: Group Work Reflection

Introduction

The group work undertaken by Liad Peretz and I set out to investigate how real-world events are reflected on Twitter. This project was driven by our shared interests in Information Engineering and our strong compatibility as partners. Our mutual curiosity greatly facilitated our project decisions, allowing us to utilize each other's strengths to effectively problem-solve throughout each project stage. Our research aimed to uncover the fascinating ways in which real-world events are mirrored in the digital realm, specifically on Twitter. This exploration not only holds academic significance but also practical relevance in the realm of Information Engineering, where understanding the dynamics of social media platforms is crucial. Our partnership was founded on the fusion of these objectives and our individual passions, setting the stage for a collaborative journey in which we each brought distinct skills and perspectives to the table.

Overview of Group Activities Throughout the Project

Throughout the project, our group's collaborative efforts led to a well-balanced distribution of tasks, fostering effective communication and coordinated time management. We approached our work with a strategic division of responsibilities for each project stage and maintained a consistent schedule of meetings, either in-person or online, tailored to the project's evolving needs. Furthermore, we conducted daily or weekly reviews of each other's work to ensure quality and progress. In addition, our weekly meetings with our supervisor played a pivotal role in obtaining key insights, valuable advice, and efficient planning for the upcoming week, which we successfully implemented. This collaborative framework not only enhanced our productivity but also contributed significantly to the project's overall success.

In the project's initial stage, a significant setback arose due to changes in Twitter's ownership structure, rendering the Twitter API inaccessible as a data source. In response to this unforeseen obstacle, I assumed the lead role in researching alternative open-source data sources and meticulously evaluating their quality, particularly with respect to the diverse event types we were investigating. This extensive research culminated in our decision to utilize Kaggle, the Internet Archive, and GDELT as our primary data sources, judiciously chosen to align with the unique requirements of each stage of our investigation. This adaptive approach allowed us to navigate and overcome the unexpected challenge, ensuring the project's continued progress and success.

Upon completing our evaluation and satisfaction with the selected data sources, we honed our focus to three specific event types: Protests, Natural Disasters, and Elections, marking the commencement of our data analysis phase. A critical aspect of this stage was the meticulous extraction and cleaning of the data.

To facilitate the setup and configuration of data extraction from the Internet Archive, GDELT, and later the Wits Cluster, Liad and I collaborated closely. We jointly deliberated on the concept, logic, environment configuration, and code structure, ensuring a well-coordinated approach. After establishing the initial infrastructure, we strategically divided the substantial workload and assigned tasks based on our strengths.

Liad took the lead in optimizing and refining the data extraction processes, ensuring the data's integrity and quality. Meanwhile, I immersed myself in the exploration of the available datasets, leveraging the collaborative features of Kaggle's Jupyter notebook environment, which we chose for its non-local, shared workspace. Within this environment, I assumed responsibility for configuring and structuring the notebook, an essential step in contextualizing each event type. This contextualization served as the foundation for unravelling the narratives concealed within the data. As part of this process, I delved deep into the data, seeking to uncover insights and patterns. Specifically investigating features such as daily or hourly tweet count ,hashtags, retweets, like counts, location metrics etc. This meticulous examination enabled us to gain a comprehensive understanding of the events under scrutiny, contributing significantly to the quality of our analysis and the depth of our findings.

Building upon our accomplishments in the extraction and analysis stage, our collaborative efforts led to the successful achievement of our primary goal: the creation of contextual patterns of interest for each event type. With this milestone behind us, we seamlessly transitioned into the next phase of our investigation, which involved raising questions about the data findings.

Firstly, we critically examined the patterns of interest we had meticulously identified and delved into understanding how these findings resonated with responses in the news media. This comparison between social media trends and news media responses was a crucial aspect of our research. Leveraging my research on GDELT and Liad's optimization of data extraction from GDELT's BigQuery platform , we successfully conducted these comparisons for each event.

Furthermore, as we delved into the analysis of the event signatures, we became curious about the nature of the trends we had uncovered. We questioned whether they were valid, indicating true representations, or anomalies. To address this, Liad focused on data extraction from the Mastodon API, a Twitter competitor. However, due to the limited user base on Mastodon, we found sufficient data only for the US Election. Concurrently, I explored alternative methods to validate these trends and opted for Google Trends, which offers both web search and news search functions. Utilizing Google Trends, I extracted search and news queries for each event, enabling a direct comparison across all events. This approach provided validation for the trends we observed in both social media and GDELT articles.

Moreover, we delved into the exploration of trends and mathematical approximations derived from the data. Guided by Dr. Martin Bekker, we explored Seasonal Decomposition and LOESS regression techniques, collaborating closely to mathematically characterize each event. During this phase, I focused on the investigation of similarity techniques, including autocorrelation and cross-correlation, while also optimizing the outcomes from the seasonal decomposition process. Simultaneously, Liad expanded data retrieval efforts for each event type, a crucial step toward creating the archetype for each event.

Ultimately, through our collaborative efforts, starting from the initial research stage and culminating in effective data synthesis, enabled us to craft a distinctive archetype for each event type.

Conclusion

In conclusion, the investigation project undertaken by Liad Peretz and I resulted in a successful investigative project. Despite initial setbacks, our robust group collaboration propelled us beyond the initial goal of analysing patterns of interest. We delved deeper into the data, ultimately establishing event archetypes for not one but three distinct events, involving an extensive volume of data exceeding one terabyte. The success of this endeavour was made possible through the effective partnership between Liad Peretz and I, as well as the invaluable guidance and support provided by our supervisor, Dr. Martin Bekker.

Appendix B: Project Plan

Appendix C: Patterns of Interest

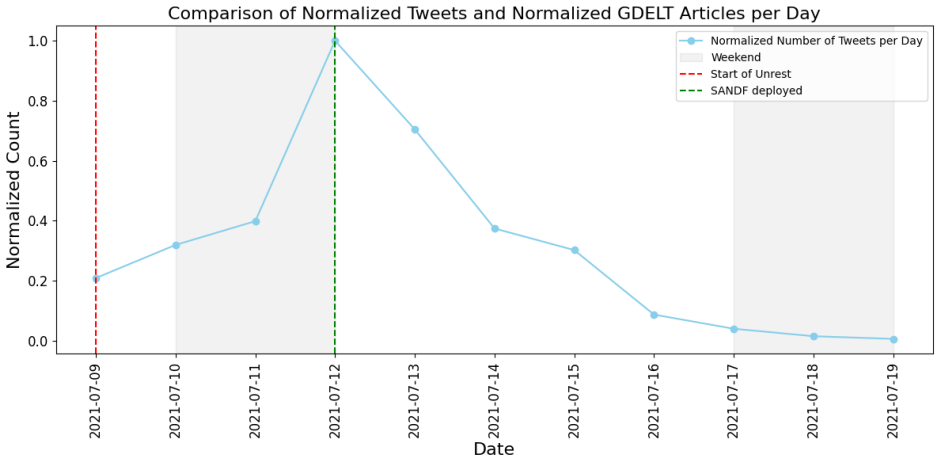

Figure 5: 2021 South African Social Unrest pattern of interest

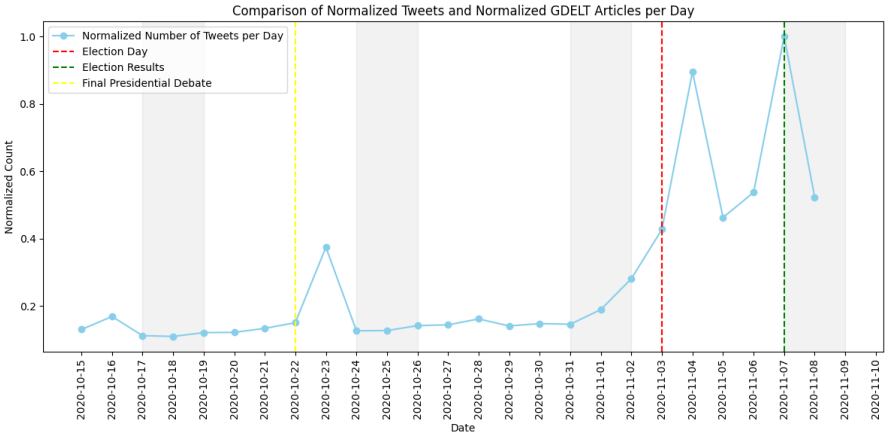

Figure 6: US Election 2020 Patter of Interest

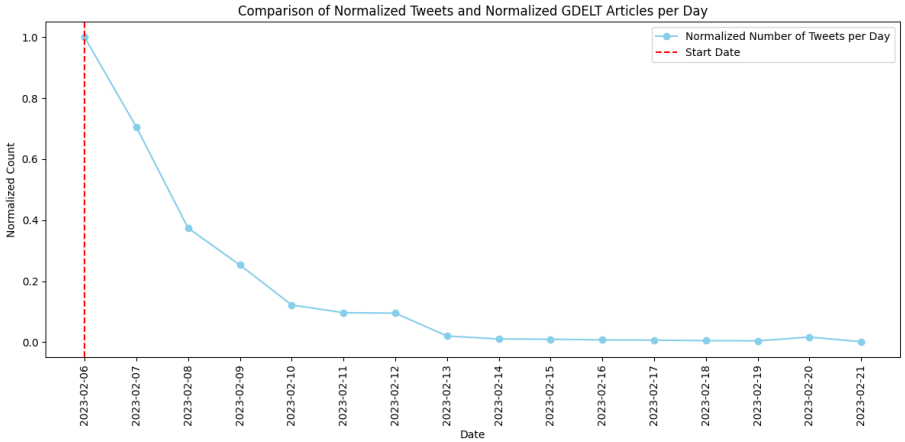

Figure 7: Turkey-Syria Pattern of Interest

Appendix D: Social Media Response compared News Media Response

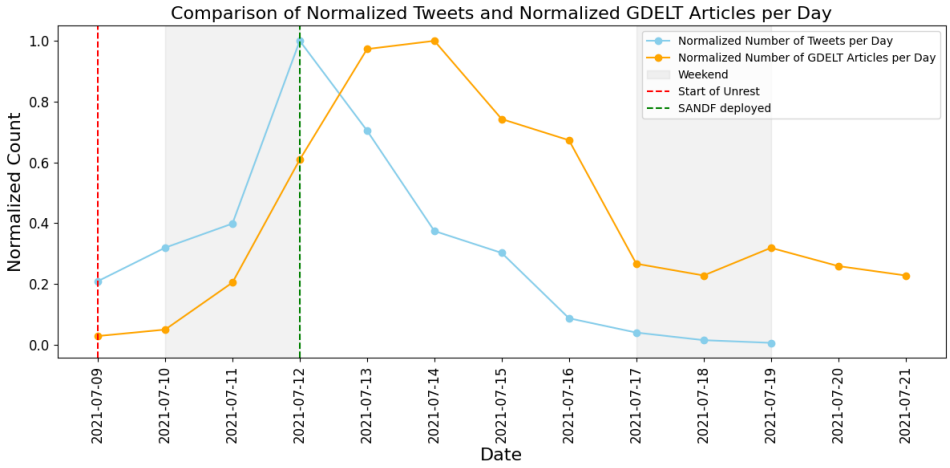

Figure 8: Comparison of social media and News Response of 2021 South African Social Unrest.

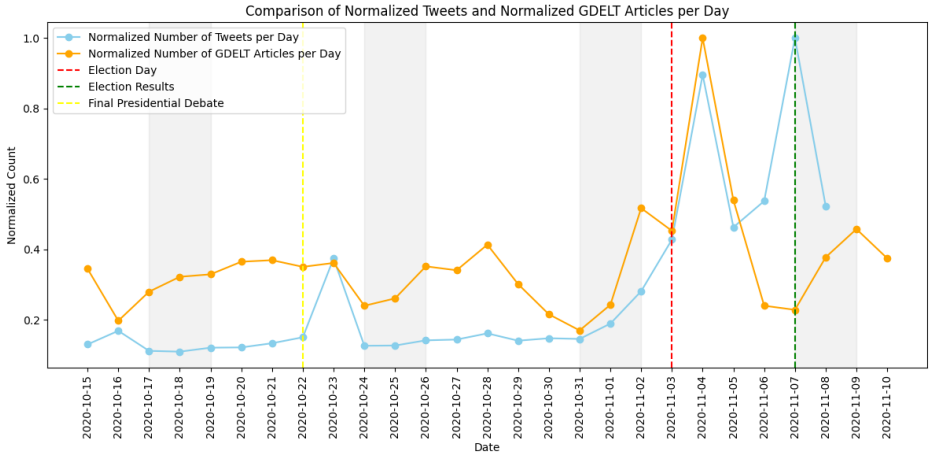

Figure 9:Comparison of social media and News Response of US Election 2020.

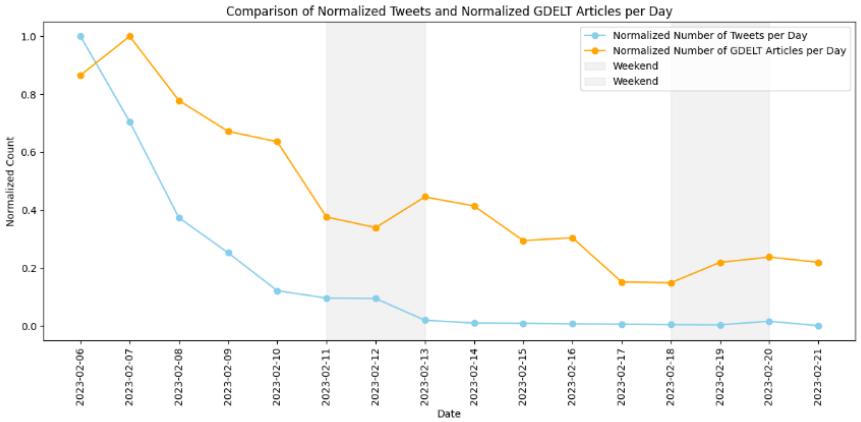

Figure 10:Comparison of social media and News Response of Turkey-Syria Earthquake

Appendix E: Trend Validation

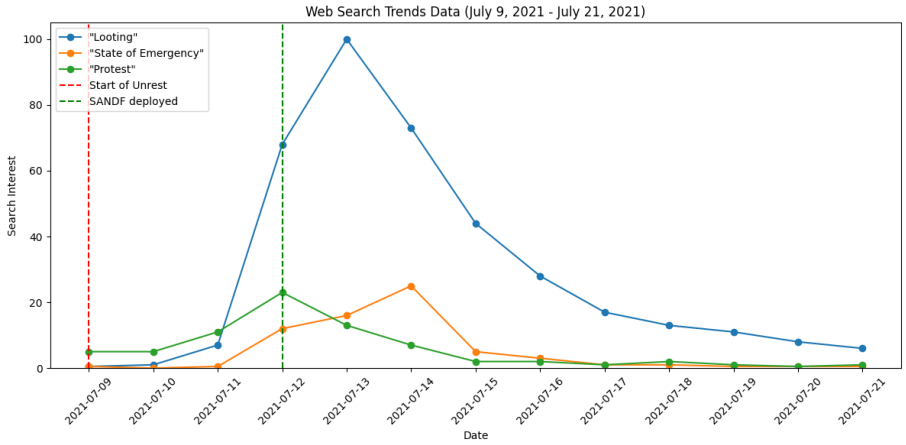

Figure 11: Trend Validation using Google Trends of 2021 South African Social Unrest

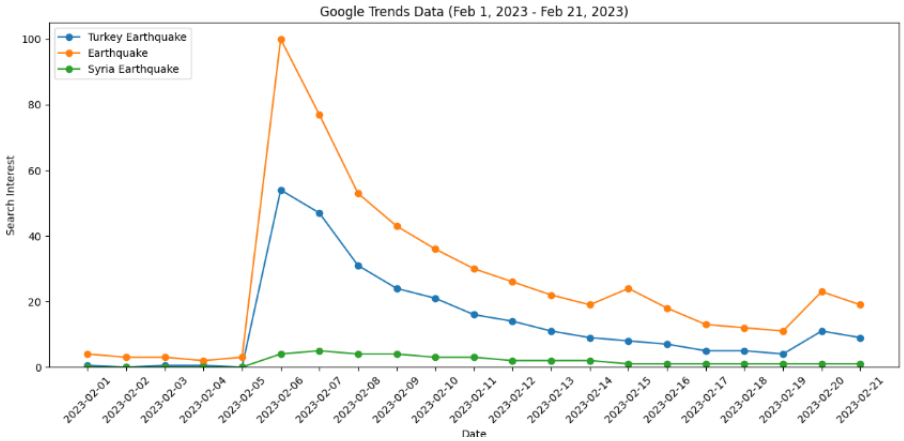

Figure 12: Trend Validation using Google Trends keywords of the Turkey-Syria Earthquake

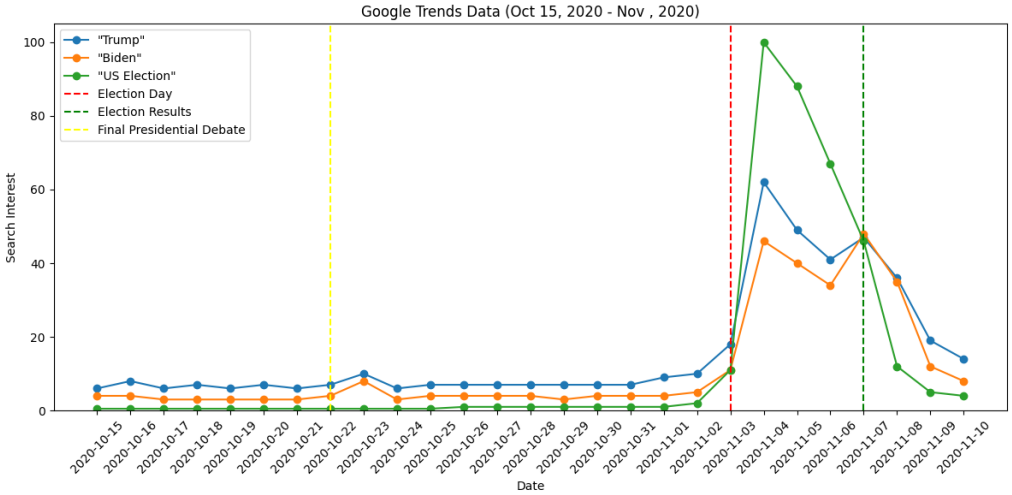

Figure 13: Trend Validation using Google Search of US 2020 Election

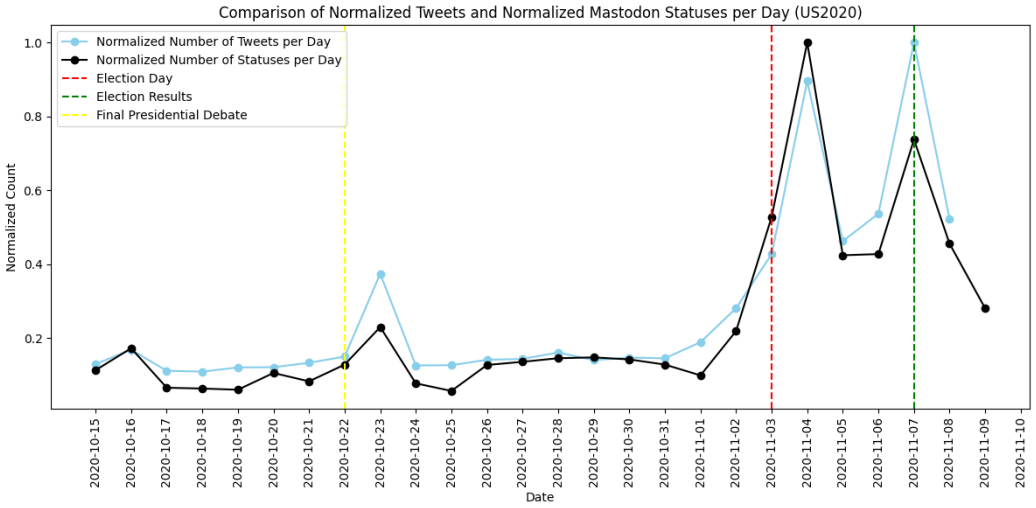

Figure 14: Trend Validation using Mastodon of the US Election in 2020

Appendix F: Event Archetypes

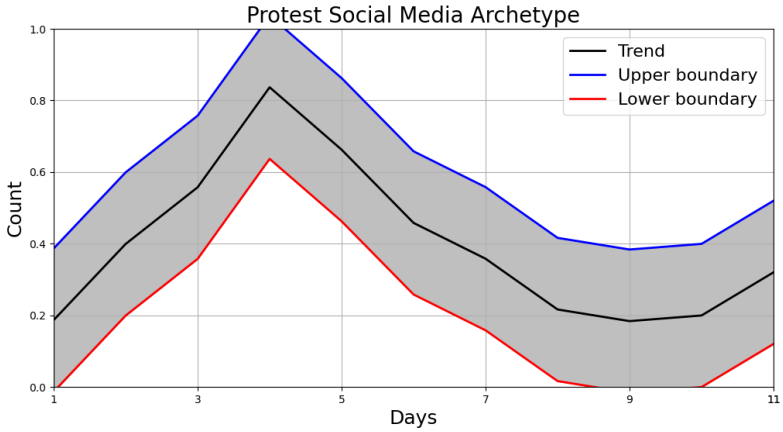

Figure 15: Event Archetype of Protest event type

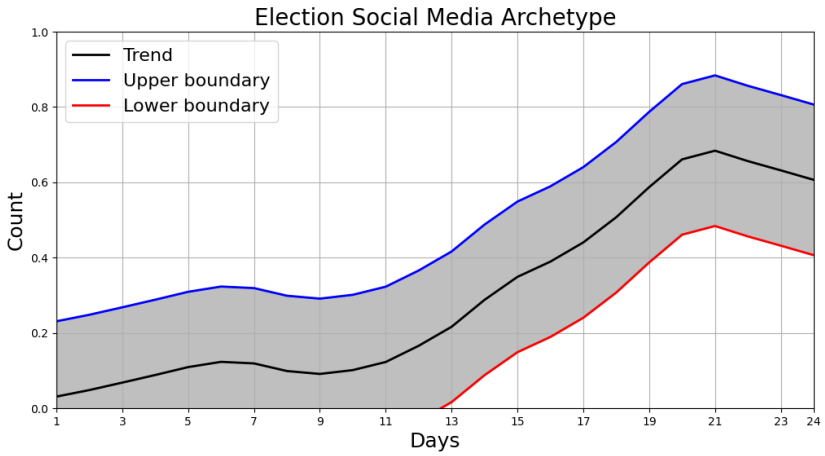

Figure 16: Event Archetype of Election event type

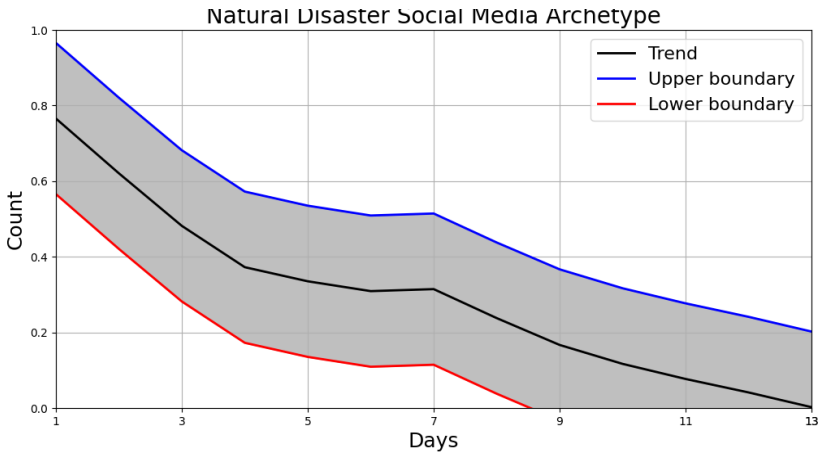

Figure 17: Event Archetype of Natural Disaster event type